A social model for Literature Access: Towards a weighted social network of authors

Lamjed Ben Jabeur, Lynda Tamine, Mohand Boughanem

A social model for Literature Access: Towards a weighted social network of authors

Lamjed Ben Jabeur
IRIT, SIG-RI
University of Paul Sabatier
118 Route de Narbonne
F-31062 Toulouse CEDEX 9
jabeur@irit.fr

Lynda Tamine
IRIT, SIG-RI
University of Paul Sabatier
118 Route de Narbonne
F-31062 Toulouse CEDEX 9
tamine@irit.fr

Mohand Boughanem
IRIT, SIG-RI
University of Paul Sabatier
F-31062 Toulouse CEDEX 9
bougha@irit.fr

ABSTRACT

This paper presents a novel retrieval approach for literature access based on social network analysis. In fact, we investigate a social model where authors represent the main entities and relationships are extracted from co-author and citation links. Moreover, we define a weighting model for social relationships which takes into account the authors positions in the social network and their mutual collaborations. Assigned weights express influence, knowledge transfer and shared interest between authors. Furthermore, we estimate document relevance by combing the document-query similarity and the document social importance derived from corresponding authors. To evaluate the effectiveness of our model, we conduct a series of experiments on a scientific document dataset that includes textual content and social data extracted from the academic social network CiteULike. Final results show that the proposed model improves the retrieval effectiveness and outperforms traditional and social information retrieval baselines.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Social search

General Terms

Formal models, Theory

Keywords

Social information retrieval, literature access, social network, social importance, social annotation

1. INTRODUCTION

Academic search engines have boosted the availability and the quality of the bibliographic resources and therefore help researchers accessing to authoritative papers and managing their references. Earlier with the introduction of the Shepard's Citations (1873) to the incoming of Google Scholar[1] (2004), scientific indexer and academic digital libraries have addressed one common issue: evaluating the importance of scientific publications. To tackle this problem, literature access used mainly bibliometrics based measures in order to estimate scientific paper quality. Afterwards, bibliographic resources are considered as hyperlink structures where citation links denote anchor text and resource authority is inferred by applying *HITS* and *PageRank* algorithms [12].

From another side, the importance of scientific publications is estimated from their social context as they are produced and consumed by social entities. This view has been dealt by Social Information Retrieval (SIR) area assuming that relevant documents are associated to important actors in the social network. In fact, the scientific importance of documents is inferred from corresponding authors centrality in the social networks which is computed using network analysis measures [20]

With this in mind and inspired by the work in [10] [15] representing bibliographic resources with a social information network, we introduce a social information retrieval model for literature access that includes additional social entities and relationships. Unlike previous work, the proposed model integrates social annotation data and information consumers in the social network and extract new relationships from citation and social bookmarking associations.

The rest of this paper is organized as follows. Section 2 reviews related work. We detail in section 3 our social information retrieval model for literature access. In section 4, we evaluate and discuss the effectiveness of our model on a dataset of scientific publications. Section 5 concludes the paper and introduces future work.

2. RELATED WORK

In the purpose of evaluating the importance of scientific publications, a wide range of researches have focused on incoming citation links as an indicator of quality and authority of scientific papers. Particularly, this feature has been used in bibliometrics in order to measure the impact of a researcher, a paper or a journal [4]. However, citation feature is not sufficient to estimate paper relevance. Therefore, measures including more factors such as link growth over time are proposed later to rank documents according to their age and

[1] http://scholar.google.com/

expected citation links [6] [14].

Other work consider the citations links as a relevance feature at both indexing and retrieval levels [18] [7]. In [18], citations improve the document descriptors (index) using terms extracted from cited documents to additionally describe the citing document. In [7], citations are viewed as hyperlinks connecting bibliographic resources in the document graph where authority is computed with the *PageRank* algorithm [17].

Despite the previously cited approaches evaluating scientific paper importance based on the citation feature, recent work use alternative measures developed in social network analysis (eg. *Betweeness* and *Closeness*...etc.) to identify central resources. These measures are applied either on the document graph [3] to highlight central documents or on the social network of authors to estimate document relevance through corresponding author's centrality [15] [9]. These works introduce the social information retrieval approaches for literature access.

Regarding the modeling of the social network for bibliographic resources, early work include only authors nodes and co-authoring relationships in the social network [21] [15]. Approaches introduced in [16] [13] extended previous binary models by assigning weights to co-author associations based on the frequency and the exclusivity of the co-authorship. Other work include documents as information nodes in the social network and align entities into document and author layers with possible associations connecting nodes from different layers [11]. In these models, relationships are extracted from the social interactions such as the collaboration, the publication and the citation.

With the introduction of the social network in the retrieval process, document relevance is not only assimilated to the query-document similarity but also interpreted as the resource authority, trust and reputation in the social network [8]. These relevance features are either modeled on the social graph as transition probabilities (integrated approach) [1] or either modeled with separate factors which are combined to estimate the final relevance of documents (modular approach) [9] [10].

In this paper we propose a social information retrieval model for literature access that estimates bibliographic resources relevance based on the social importance of their authors. Unlike related work and our previous contribution [19], this model has several new features:

- First, the social information network includes new entities corresponding to users and social annotations in addition to documents and author nodes presented in [11] [9]. This helps to estimate document relevance based on thier social production and consuming contexts.
- Second, we include citations links as social interactions between authors of scientific papers enriching thus their mutual associations previously based on co-author relationships only [10] [15].
- Finally, we define a weighting model for edges connecting social entities in the contrast of approaches in [9] [15] modeling bibliographic resources by a binary network model. Specifically, weights are assigned to co-authorship, citation and authorship edges in order to evaluate influence, knowledge transfer and shared interest between authors.

3. THE SOCIAL INFORMATION RETRIEVAL MODEL

An information retrieval model is a theoretical support that aims at representing documents and queries and measuring their similarity viewed as relevance. Formally and based on the representation introduced in [2], the social information retrieval model can be represented by a quintuple $[D, Q, G, F, R(q_i, d_j, G)]$ where D is the set of documents, Q represents the set of queries, G is the social information network, F represents the modeling process of documents and queries and $R(q_i, d_j, G)$ is the ranking function including various social relevance features and taking into account the social information network topology G. This function can be defined by combining the subset of the flowing factors: the topical relevance, the social importance of actors, the social distance, the popularity, the freshness and the incoming links and tags [1].

The social information network G represents the social entities that interact in the social producing and consuming context of documents. As illustrated in figure 1, the social information network G include all actors and the data that help to estimate the social relevance of documents. In fact, actors represent information producers (authors) and information consumers (users) whereas data cover documents and social annotations (tags, rating, reviews). Accordingly, actors become information nodes collaborating to produce documents and interacting to provide social annotations.

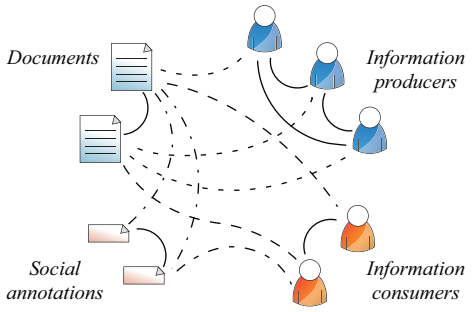

Figure 1: The Social Information Network

The social information model can be represented by a graph $G = (V, E)$ where nodes $V = A \cup U \cup D \cup T$ denote social entities with A, U, D and T respectively correspond to authors, users, documents and social annotations. The set of edges $E \subseteq V \times V$ represents social relationships connecting various node types (authorship, co-authorship, friendship, citation, annotation...etc). Within this generic model, we present in what follows the social information network model for bibliographic resources and then we detail our weighting schema for social relationships.

3.1 A Social Information Network for Literature Access

Previous work [11] [10] model the social network of academic resources using only information producing context. However with the introduction of the social network of scientists on the web (e.g. CiteULike[2] and Academia[3]) users participate also to provide additional descriptors for bibliographic resources. Unlike the friends-of-friends social applications such as FaceBook[4] and MySpace[5], academic social networks may express specific relationships between social entities. We identify the following social relationships that are involved with documents, authors, users and tags nodes:

- **Authorship:** connects an author $a_i \in A$ with his authored document $d_j \in D$.
- **Reference:** connects a document $d_i \in D$ with its referenced documents.
- **Co-authorship:** connects two authors $a_i, a_j \in A$ having produced one common document at least
- **Citation:** connects two authors $a_i, a_j \in A$ with author a_j is cited by a_i at least once through his documents.
- **bookmarking:** connects a user $u_i \in U$ and his bookmarked document $d_j \in D$.
- **Annotation:** connects a document $d_i \in D$ with a tag $t_j \in T$ assigned at least once to describe its content.
- **Tagging:** connects a user $u_i \in U$ and a tag $t_j \in T$ as he use it at least once to bookmark a document.
- **Friendship:** connects users $u_i, u_j \in U$ if either they have a direct personal relationship or they join the same group.

The social entities included in the social information network for the literature access could be represented using a graph notation illustrated in figure 2.

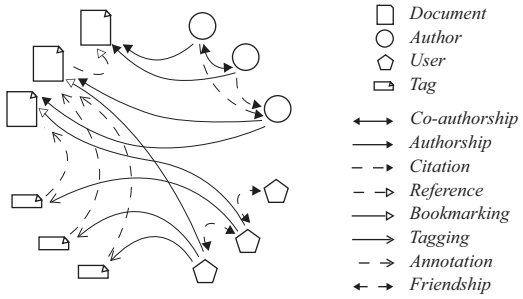

Figure 2: The Social Information Network for Literature Access

3.2 Weighting the social relationships

Edges connecting social nodes may express different kinds of social relationships and significantly optimize the exploring process of the social network. In fact, if we explore the social neighborhood (closely reachable nodes) of a social entity, weights would help to select jumping nodes. In this work, we are mainly interested in the social network of scientific publications and we investigate a weighting model for author-to-author relationships $e(a_i, a_j) \in (A \times A)$ and author-to-document relationships $e(a_i, d_j) \in (A \times D)$.

a. Co-authorship: This social relationship is represented by an undirected edge connecting two authors having collaborated to produce a document. Co-authors have often personnel direct relationships however multiple collaborations reflect their similarity and shared interest. In fact, scientific authors tend towards exchanging knowledge and diversifying their collaborations. For this reason, we propose to normalize weights by the total of collaborations involving the couple of authors. The co-authorship edges could be weighted as follows:

$$Co(i,j) = \frac{2A(i,j)}{A(i) + A(j)} \quad (1)$$

Where $A(i)$ is the number of documents authored by a_i and $A(i,j)$ represents the number of documents co-authored by a_i and a_j.

b. Citation: This social relationship is represented by a directed edge connecting an author with his cited authors. An author who usually cites a second author would be influenced by his opinions and eventually discuss similar subjects. Therefore, the citation links expresses knowledge transfer between authors. To evaluate citation relationship strength, we propose to take into consideration the citation frequency as well as the total announced citations. Citation relationships is weighted as follows:

$$Ci(i,j) = \frac{C(i,j)}{C(j)} \quad (2)$$

Where $C(i)$ is the number of citations announced by author a_i and $C(i,j)$ represents the number of times author a_i cites a_j.

c. Authorship This social relationship is represented by a directed edge connecting an author with his authored documents. The strength of the authorship association is viewed as the author affiliation to the topic of the document. We note that an author would be more affiliated to a topic if he frequently addressed in his published papers. Therefore, a co-author will be more associated to a document d discussing a topic S rather than all his co-authors if he has published more documents in this topic.

In order to estimate the knowledge and the experience of a co-author $a_k \in A$ on the topic of document d, we propose to compare the quantity of information he has imported via his other publications. From the information producer point of view, this can be measured by the information entropy $H_d^k(t_i)$ for tags assigned to the sub-set of the co-author publications noted $\mathcal{A}_k$. We consider as a random variable each tag $t_i \in T^d$ assigned to document d where it exists an edge $e(t_i, d_j) \in (T \times D)$ and we calculate its probability distribution $Pr^k(t_i)$ among the sub-collection of documents $\mathcal{A} = \bigcup_{k=1}^{m} \mathcal{A}_k$ published by the m co-authors of the document d.

[2]http://www.citeulike.org/
[3]http://www.academia.edu/
[4]http://www.facebook.com
[5]http://www.myspace.com

We propose to normalize the information entropy values by the number of tags associated to the document noted $\|T^d\|$. Meanwhile, a co-author with a single publication in the collection gets a higher weight value $w(a_i, dj) = 1$ rather than his co-authors with much more publications on the topic of the document. We propose so to assign a default weight value for authors having unique document in the dataset and take into consideration the number of publications per author. The final *Authorship* weight $w(a_k, d)$ is computed as follows:

$$w(a_k, d) = \left[1 - \frac{1}{\|T^d\|} H_d^k(t_i)\right] - \frac{1}{\|\mathcal{A}_k\|}\theta \quad (3)$$

where

$$H_d^k(t_i) = -\sum_{t_i \in T^d} Pr^k(t_i) log Pr^k(t_i) \quad (4)$$

$$Pr^k(t_i) = 0,5\frac{tf(t_i, \mathcal{A}_k)}{tf(t_i, \mathcal{A})} + 0,5 \quad (5)$$

With $tf(t_i, \mathcal{A}_k)$ is the frequency of tag t_i in the subset of author a_k documents ($\mathcal{A}_k$) and $tf(t_i, \mathcal{A})$ represents the tag frequency in the sub-collection of the co-authors documents ($\mathcal{A}$). In order to get ascendant values of entropy, we scale tag probability into the interval $[0.5, 1]$. We note that $1 - \theta$ is the default weight value attributed to authors having a single document in the collection.

Some social network analysis algorithms do not support multiple edges between two nodes with similar directions. Thus we propose to combine the co-authorship and citation weights as follows:

$$w(i, j) = \frac{1}{4} * (1 + Co(i, j)) * (1 + Ci(i, j)) \quad (6)$$

3.3 Computing Social Relevance

The idea of document relevance estimation within the social network is to derive a more accurate response to the user by combining the topical relevance of document d and the importance of associated authors in the social network.

In this work, we aim to select the social importance measures that identify central authors in the social network of bibliographic resources. Therefore, we compute for each author a social importance score $C_G(a_i)$ using one of the following importance measures: the *Betweeness*, the *Closeness*, the *PageRank*, the *HITS*'s *Authority* score and the *HITS*'s *Hub* score. We apply these importance measures only on the sub-graph of authors $G_a = (A, E_a)$ where $E_a \subseteq (A \times A)$ and edges denote either co-authorships or either citation links and weighted as described previously.

Afterwards, a social importance score is transposed from authors to documents using a weighted sum aggregation as follows:

$$Imp_G(d) = \sum_{i=1}^{k} w(a_i, d) C_G(a_i) \quad (7)$$

The social score of document $Imp_G(d)$ estimates its social relevance. Nevertheless, it's not enough to retrieve relevant documents from the collection. Therefore, we combine $Imp_G(d)$ score with a traditional information retrieval metric such a TF-IDF score as follows:

$$Rel(d) = \alpha RSV(q, d) + (1 - \alpha) Imp_G(d) \quad (8)$$

Where $\alpha \in [0..1]$ is a weighting parameter, $RSV(q, d)$ is a normalized similarity measure between query q and document d, $Imp_G(d)$ is the importance of document d in the social network G.

4. EXPERIMENTAL EVALUATION

In order to evaluate the effectiveness of our model, we conduct a series of experiments on a scientific documents dataset published on the ACM SIGIR conference from 1978 to 2008. The main evaluation objectives are:

- Comparing different importance measures with both binary and weighted social network models to estimate scientific papers importance.
- Evaluating the effectiveness of our model compared to traditional information retrieval models and other closely related retrieval models.

4.1 Experimental datasets and design

We used for experiments the SIGIR dataset that contains informations about authors and citation links in addition to the textual content of publications. We included in the social network all authors having published at least one paper in the ACM SIGIR conference. Two authors are associated with a social relationship if either they co-authored a SIGIR publication or one of them cites the other author through his SIGIR paper.

To enrich this dataset, we gathered data about information consumers and social interactions from the academic social network CiteULike. We collected all social bookmarks targeting the SIGIR publications and we extracted related tags and corresponding users.

The following paragraphs describe the dataset characteristics and evaluation measures.

- **Social network properties:** The SIGIR dataset includes 2871 authors with an average of 2 co-authorships and 16 citation links per author. As shown in table 1, the citation relationships dominate the social network with 9 times much more edges than co-authorship associations. In fact, including citation links restructure small and dispersed components into larger author communities. Consequently, the giant component connecting the majority of authors nodes is enlarged with citation relationships to include 84% of authors as shown is figure 3.

- **Queries and relevance assumption:** Tags are user generated keywords in order to annotate document content. They help user indexing document from their point of view and consequently correspond to a later information needs possibly satisfied with this document. Unlike automatic extracted terms form textual context, tags seem to be more convenient to represent

Authors	2871
Co-authorships	5047
Citation links	45880
Co-authorship and/or citation links	52516

Table 1: Social network properties of the SIGIR dataset

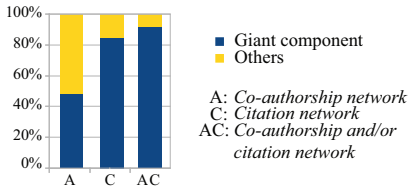

Figure 3: The giant component of the SIGIR social network

queries once both of them are user-generated terms expressing information needs. Thus we propose to choose tags assigned to the SIGIR publications as representative queries in our experiments.

We assume that the popular tags are more important in the social context. Thus we select as queries the most frequent tags assigned to the SIGIR publications then we build the ground truth as follows:

Step 1: We select as initial queries the top 100 tags sorted by total bookmarks targeting the SIGIR publications (popular tags).

Step 2: We remove personal and empty tags such as *"to read"* and *"sigir"*.

Step 3: We regroup similar tags with different forms like *"language model"* and *"language modelling"*.

Step 4: For each query, we collect documents bookmarked at least once by the corresponding tag or its similar forms.

Step 5: From the previous list of documents corresponding to a query tag, we select only the documents having the query tag among their 3 top assigned tags. The final document set corresponds to the query relevant documents.

Step 6: We remove the query tag if no relevant document is found.

We retain for experimentation the top 25 queries and their corresponding relevant documents. The final collection includes 512 relevant documents with an average of 20 relevant documents per query. To index the dataset, we used the open source library for information retrieval Apache Lucene[6] which is based on a modified scoring function of the vector-space model described in [5].

- Evaluation measures: In oder to compare the social importance measures and evaluate our model performance, we use recall and precision. Users are commonly interested in the top results, therefore we study precision at 0.1 and 0.2 points of recall. With an average of 800 retrieved documents per query, these recall points correspond to the 160 first documents.

[6]http://lucene.apache.org/

4.2 Comparison of social importance measures

The social importance measures highlight key entities in the social network and include measures introduced by both domains of social network analysis [20] and hyperlink analysis [17] [12]. These measures have multiple semantics which vary from one social application to another. In the context of scientific publications, the *Betweeness* measure is considered as an indicator of interdisciplinarity and highlights authors connecting dispersed partitions of the scientific community. The *Closeness* measure, based on the shortest path in the graph, reflects the reachability and independence of an author in his social neighborhood. The *PageRank* measure and the *Authority* score computed by *HITS* algorithm distinguish the authoritative resources in the social network. In contrast, the *Hub* score computed by *HITS* algorithm identifies authors having an important social activity and relying on authoritative resources, these authors are called *Centrals*.

Table 2 lists the top 10 authors ranked via the following social importance measures: the *Betweeness*, the *Closeness*, the *PageRank*, the *Authority* score and the *Hub* score. We note that some authors such as *W. B. Croft* figure at the top of many lists and this is due to their strategic positions in the social network. In contrast, some authors, such as *S. Dumais* in *Betweeness* list and *E. Kokiopoulou* listed in the top *Closeness* ranking, figure only at one list and this is due to their collaborations relying either on close or disparate communities.

We applied the social importance measures listed above on both a binary and a weighted model of the social network. We note *W-Betwenness* the application of *Betwenness* measure on the weighted model of the social network. We use the same notation for the rest of social importance measures.

Table 3 presents comparative effectiveness results of the different importance measures for both binary and weighted models of the social network. These results are obtained using only the social importance score of documents by setting $\alpha = 0$ in formula 8. We note that the *Hub* measure better ranks scientific papers for both binary and weighted models of the social network. We conclude that the importance of scientific publications can be estimated as the *Centrality* of their authors.

The weighted model slightly improves the retrieval precision for most social importance measures. This is approved with values obtained by *W-Hub*, *W-Authority* and *W-Betweens* measures beyond their analogous measures applied to a binary social network. Therefore, we conclude that the properties expressed through weights on social relationships including the shared interests, the influence and the knowledge transfer can better identify *Central* authors and then estimate relevance of bibliographic resources.

For all social importance measures, precisions $p@0.1$ and $p@0.2$ not exceed the threshold of 60% compared to those of the traditional information retrieval model based on the TF*IDF metric having $p@0.1 = 0.08$ and $p@0.2 = 0.0786$. Therefore, the social importance measures are not able to sort the results without taking into account the similarity between document and query.

Betweeness	Closeness	PageRank	Authority	Hub
W. B. Croft	E. Kokiopoulou	W. B. Croft	W. B. Croft	W. B. Croft
J. Allan	Y. Saad	C. Buckley	C. Buckley	C. Zhai
C. Zhai	M. Zhong	E. M. Voorhees	E. M. Voorhees	J. Zobel
M. Sanderson	X. Huang	S. Robertson	J. Xu	W. Ma
S. Dumais	U. Deppisch	C. J. V. Rijsbergen	C. Zhai	C. L. A. Clarke
J. Zobel	D. D. Jaco	J. Allan	J. Allan	J. Allan
C. Buckley	G. Garbolino	D. D. Lewis	A. Singhal	J. Callan
S. Robertson	M. W. Davis	N. J. Belkin	S. Robertson	J. Wen
C. Yu	W. C. Ogden	J. O. Pedersen	J. Lafferty	O. Frieder
J. Nie	A. Stent	G. Salton	M. Mitra	J. Nie

Table 2: Top 10 authors ranked by social importance measures

In the remaining experiments, we retained the *W-Hub* measure as it is the best measure expressing the social importance of bibliographic resources.

	p@0.1	p@0.2		p@0.1	p@0.2
Betweeness	0,0363	0,0363	W-Betweeness	0,0374	0,0398
Closeness	0,0232	0,0191	W-Closeness	0,0214	0,0189
PageRank	0,0324	0,0299	W-PageRank	0,0225	0,0199
Authority	0,0389	0,0411	W-Authority	0,0398	0,0423
Hub	**0,0516**	**0,0430**	W-Hub	**0,0516**	**0,0433**

Table 3: Comparison of social importance measures

4.3 Evaluation of our model effectiveness

To evaluate the effectiveness of our model, we first select the best tuning parameter α then we compare the retrieval performances with similar retrieval systems.

4.3.1 Tuning the parameter α

We studied the impact of the parameter α on the retrieval process (see formula 8). We note that if $\alpha = 0$ only the social relevance is taken into account. Moreover, $\alpha = 1$ corresponds to the baseline TF * IDF since the topical relevance is only considered to rank documents.

We note through figure 4 a significant improvement in performance following the integration of topical relevance with a value of α over 0.4. Analyzing precisions p@0.1 and p@0.2 depending on the parameter α shows that the curves have peaks whose values exceed the value obtained for $\alpha = 1$, and that when the topic relevance is only taken into consideration. So the combination of the two scores can effectively improve the final ranking of documents. The best values of the parameter α is obtained between 0.5 and 0.6.

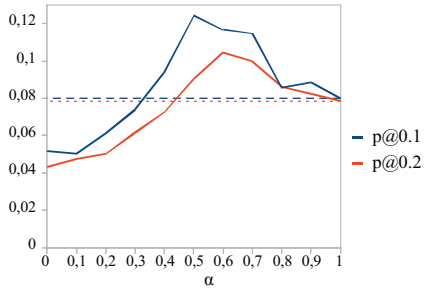

Figure 4: Tuning the α parameter

4.3.2 Performance evaluation

We compare our model with 3 baselines detailed as follows:

- **TF*IDF model :** denotes a traditional information retrieval system implemented by Apache Lucene based on the TF*IDF metric and using the stemming algorithm *SnowBall Stemmer*. We used this retrieval system with the same configuration in our model to select documents and compute their topical relevance.
- **PR-Docs model:** denotes a retrieval system that estimates the importance of documents based on their authority. It combines the topical relevance and the *PageRank* score of documents computed on the document graph where edges represent citation links. Final document relevance is computed as fallows:

$$Rel(d) = \alpha RSV(q,d) + (1-\alpha) PageRank_{docs}(d) \quad (9)$$

We note that the topical relevance $RSV(q,d)$ is computed using the first baseline TF*IDF. We studied the impact of the α parameter on the search effectiveness and we note that best retrieval precisions are obtained with $\alpha = 0.7$ for p@0.1 and $\alpha = 0.3$ for p@0.2 as shown in figure 5.

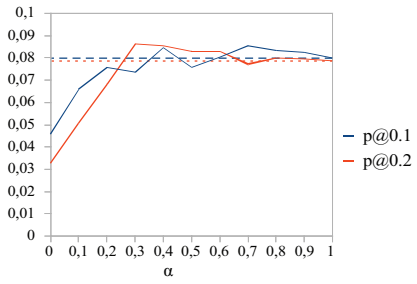

Figure 5: Tuning the α parameter for PR-Docs model

- **Kirsch's model:** denotes the social information retrieval model introduced in [10] that represents authors using a binary co-authorship network and computes their social importance score using the *PageRank* measure. This model combines the topical relevance and the social relevance as follows:

$$Rel(d) = RSV(q,d) * r_d \quad (10)$$

with r_d is the social relevance of the document d computed as the sum of its authors *PageRank* scores.

Figure 6 compare results obtained by the different baselines and our social model tuned with $\alpha = 0.5$ and $\alpha = 0.6$ noted respectively $SM0.5$ and $SM0.6$. We note that the best values of the parameter α can lead to an improvement in favor of our model between 15% and 55% compared to the baseline TF*IDF. Therefore, we confirm that integrating the social relavance of document can significantly imporove the retrieval effectiveness.

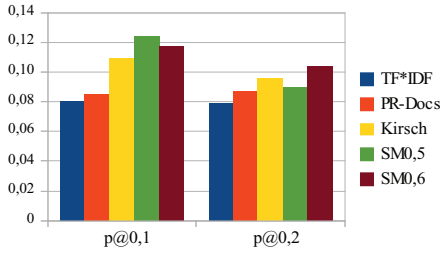

Figure 6: Evaluation of the retrieval effectiveness

Comparing our model to best obtained values of the PR-Docs model, we note an improvement of 45% for p@0.1. Therefore, we conclude that the *W-Hub* measure computed on the social network of authors expresses better the social importance of scientific papers than prior measures based on the citation graph.

Comparing our model to the Kirsch's model, we note an improvement of 14% that confirms the impact of including the citation links and weighting the social network edges on the retrieval performances.

In summary, results show a low performance of compared retrieval systems. In fact we used tags for experimental evaluation which are user generated-terms and may not be present in document content. Therefore, only a few relevant documents can be retrieved which explains the low precisions of proposed model.

Furthermore, results are proportional to the content-based retrieval model used to compute the topical relevance of documents and its performance directly affects the extended models. The main objective of previous experiments is to ameliorate content-based raking by including the social importance of document and this is achieved with significant improvement of 55% compared to the TF*ID baseline.

5. CONCLUSION

We proposed in this paper a social information retrieval model for literature access. Our model includes new social relationships in the social information network such as citation links and social annotation associations and a weighting schema for social edges. Our experimental evaluation on the SIGIR dataset reveals that the *Hub* measure is able to better evaluate the importance of scientific documents and shows the superiority of our model with a significant rate compared to traditional information retrieval model and other closely related models.

In future works we plan to integrate more social relevance features such as the social distance between the querier and retrieved documents. We plan also to conduct experiments on a scientific document dataset that covers various research areas.

6. REFERENCES

[1] S. Amer-Yahia, M. Benedikt, and P. Bohannon. Challenges in searching online communities. *IEEE Data Eng. Bull.*, 30(2):23–31, 2007.

[2] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. ACM Press, New York, 1999.

[3] J. Bollen, H. V. de Sompel, J. A. Smith, and R. Luce. Toward alternative metrics of journal impact: A comparison of download and citation data. *CoRR*, abs/cs/0503007, 2005.

[4] E. Garfield. The history and meaning of the journal impact factor. *JAMA*, 295(1):90–93, 2006.

[5] E. Hatcher and O. Gospodnetic. *Lucene in Action (In Action series)*. Manning Publications Co., Greenwich, CT, USA, 2004.

[6] C. Hauff and L. Azzopardi. Age dependent document priors in link structure analysis. In *ECIR*, pages 552–554, 2005.

[7] D. Hawking and N. Craswell. The very large collection and web tracks. In E. Voorhees and D. Harman, editors, *TREC: Experiment and Evaluation in Information Retrieval*. MIT Press, 2005.

[8] G. Kazai and N. Milic-Frayling. Trust, authority and popularity in social information retrieval. In *CIKM '08*, pages 1503–1504, New York, NY, USA, 2008. ACM.

[9] L. Kirchhoff, K. Stanoevska-Slabeva, T. Nicolai, and M. Fleck. Using social network analysis to enhance information retrieval systems. In *in Applications of Social Network Analysis (ASNA) (Zurich) , 12-9-2008*, 2008.

[10] S. M. Kirsch, M. Gnasa, and A. B. Cremers. Beyond the web: Retrieval in social information spaces. In *In Proceedings of the 28 th European Conference on Information Retrieval (ECIR 2006*. Springer, 2006.

[11] N. T. Korfiatis, M. Poulos, and G. Bokos. Evaluating authoritative sources using social networks: an insight from wikipedia. *Online Information Review*, 30(3):252–262, 2006.

[12] A. N. Langville and C. D. Meyer. A survey of eigenvector methods for web information retrieval. *SIAM Rev.*, 47(1):135–161, 2005.

[13] X. Liu, J. Bollen, M. L. Nelson, and H. V. de Sompel. Co-authorship networks in the digital library research community. *CoRR*, abs/cs/0502056, 2005.

[14] E. Meij and M. de Rijke. Using prior information derived from citations in literature search. In *RIAO*, 2007.

[15] P. Mutschke. Enhancing information retrieval in federated bibliographic data sources using author network based stratagems. In *Reserach and Advanced Technology for Digital Libraries: 5th European Conference, ECDL 2001, Darmstadt, Germany, September 4-9, 2001; Proceedings*, 2001.

[16] M. E. J. Newman. Who is the best connected scientist? a study of scientific coauthorship networks. Working Papers 00-12-064, Santa Fe Institute, Dec

2000.

[17] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project, 1998.

[18] A. Ritchie, S. Teufel, and S. Robertson. Using terms from citations for ir: Some first results. In *ECIR*, pages 211–221, 2008.

[19] L. Tamine, A. B. Jabeur, and W. Bahsoun. An exploratory study on using social information networks for flexible literature access. In *FQAS*, pages 88–98, 2009.

[20] S. Wasserman, K. Faust, and D. Iacobucci. *Social Network Analysis : Methods and Applications (Structural Analysis in the Social Sciences)*. Cambridge University Press, November 1994.

[21] E. Yan and Y. Ding. Applying centrality measures to impact analysis: A coauthorship network analysis. *J. Am. Soc. Inf. Sci. Technol.*, 60(10):2107–2118, 2009.